ファイルの最適バックアップ戦略について (決定理論とその関連分野)

AUTHOR(S):

濱田, 年男

CITATION:

濱田, 年男. ファイルの最適バックアップ戦略について (決定理論とその関連分野). 数理解析研究所講究録 1998, 1043: 67-75

ISSUE DATE:

1998-04

URL:

http://hdl.handle.net/2433/62130

数理解析研究所講究録
1043巻 1998年 67-75

ファイルの最適バックアップ戦略について

神戸商科大学　濱田年男　(Toshio Hamada)

1　緒言

コンピュータやワードプロセッサーを用いてフロッピーディスク上にファイルを作成する場合に，機器の操作ミスや電源の異常，ディスクの破損や紛失等，種々の理由により発生したトラブルによって，ファイルを読み出せなくなる場合が生じる．このようなトラブルに対処するために，通常はバックアップ用のフロッピーディスクにファイルをコピーして保存し，現在使用中のフロッピーディスクに異常が生じた場合に，バックアップ用のフロッピーディスクを用いて壊れたフロッピーディスクを復元することができる．しかし，ファイルのバックアップを取ることは，手間がかかるので，ファイルを1つ作成するごとにバックアップは取らず，いくつかのファイルを作った後で，まとめてバックアップを取ることがある．どこでバックアップを取るべきかは重要な問題である．このような問題に関連した研究としては，[3],[4],[6],[7],[9],[10],[11] を始め多くの研究がなされてきている．

本研究では，フロッピーディスクの容量に比べて，小さなファイルを多数作成する場合を想定し，いくつのファイルを作成したときに，バックアップを取るのが最適であるかを考えるための数学的モデルを構築し，最適なバックアップの方法を考える．

2　モデルと定式化

図1のように，現在コンピュータを用いてフロッピーディスク D_1 上で作業を行っているものとする．このフロッピーディスク D_1 の中にバックアップを取っていない k 個のファイルが存在し，あと残り n 個のファイルを新たに作成する予定であり，これら n 個の中の最初のファイルをちょうど作成し，保存し終えたという時点を考える．この状態を (n,k) で表すことにする．このとき，このフロッピーディスク D_1 内に，別のフロッピーディスク D_2 内にコピーが存在する l 個のファイルが存在しても，これらは仮に消滅した場合には，再生することが可能であり，従ってこれらの存在は無視することができるものとする．

状態 (n,k) においてバックアップを取るか取らないかのいずれかの決定を選ぶことができるものとする．バックアップを取るという決定を a_0，取らないという決定を a_1 で表すことにする．a_0 を選択すると，状態は確率1で $(n-1,0)$ に推移し，このときバックアップに要する費用（手間等も含めて考える）として C がかかり，また $k+1$ 個のファイルはバックアップを取られたので，もはや安全であるとみなし，ファイル1個あたり R の利得が生じるものとする．また，a_1 を選択すると，バックアップは作成しないので，次の決定時点までにファイルが消滅する可能性がある．このようなトラブルが生じる確率を p とすると，トラブルが生じたら状態は $(n-1,0)$ に推移し，消滅したファイル1個あたりの損失が B であるとし，トラブルが生じなければ状態は $(n-1,k+1)$ に推移する．

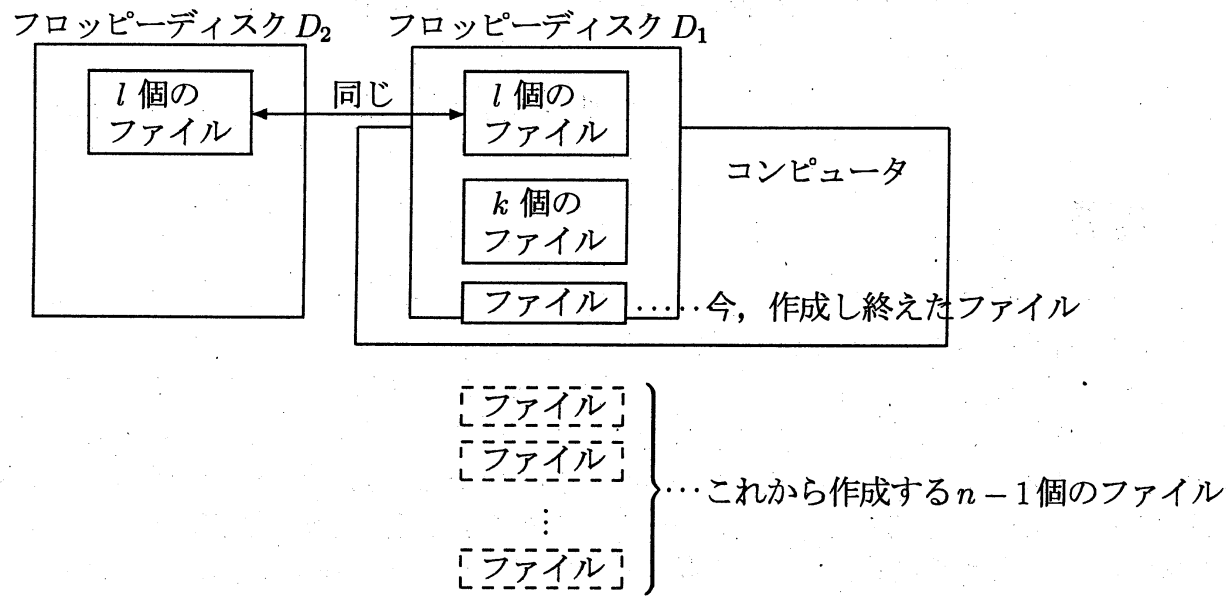

図1. (n,k) で表される状態

状態 (n,k) において，以後最適政策を用いたときの最大期待総利得を $f_n(k)$ とおき，また状態 (n,k) において，まず $a_i(i=0,1)$ を行い，以後最適戦略を用いたときの最大期待総利得を $f_n^i(k)$ で表すことにする．このとき，$n=1,2,3,\cdots$ および $k=0,1,2,\cdots$ に対して

$$f_n(k)=\max\left\{f_n^0(k),f_n^1(k)\right\} \tag{1}$$

であり

$$f_0(k)=kR \tag{2}$$

となる．ここに

$$f_n^0(k)=-C+(k+1)R+f_{n-1}(0) \tag{3}$$

および

$$f_n^1(k)=p\{-(k+1)B+f_{n-1}(0)\}+(1-p)f_{n-1}(k+1). \tag{4}$$

である．ここで，もし状態 (n,k) において $f_n^i(k)\geq f_n^{1-i}(k)$ ならば，$a_i(i=0,1)$ が最適である．状態 (n,k) において $f_n^0(k)=f_n^1(k)$ ならば，a_0 と a_1 の両方が最適であるが，便宜上 a_0 が最適であるということにする．このとき，n についての帰納法により，

$$-B\leq f_n(k)-f_n(k-1)\leq R \tag{5}$$

および

$$-kB\leq f_n(k)-f_n(0)\leq kR. \tag{6}$$

が導ける.

ここで, $n=1,2,3,\cdots$ および $k=0,1,2,\cdots$に対して.

$$d_n(k)=f_n^0(k)-f_n^1(k) \tag{7}$$

と定義すると, (3) と (4) により

$$d_n(k)=-C+(k+1)(R+pB)+(1-p)\{f_{n-1}(0)-f_{n-1}(k+1)\}.$$

および

$$d_n(k-1)=-C+k(R+pB)+(1-p)\{f_{n-1}(0)-f_{n-1}(k)\}.$$

が得られる. したがって

$$d_n(k)-d_n(k-1)=(R+pB)-(1-p)\{f_{n-1}(k+1)-f_{n-1}(k)\}.$$

となる. ここで(5) により

$$-B\leq f_{n-1}(k+1)-f_{n-1}(k)\leq R$$

だから

$$d_n(k-1)+p(B+R)\leq d_n(k).$$

となり, この不等式により次の定理が得られる.

定理 1 (*i*) 状態 $(n,k-1)$ においてバックアップを取るのが最適ならば, 状態 (n,k) においてもバックアップを取るのが最適である.
(*ii*) 状態 (n,k) においてバックアップを取るのが最適であっても,

$$p(B+R)\geq d_n(k)$$

ならば, 状態 $(n,k-1)$ においてバックアップを取らないのが最適である.
(*iii*) 状態 $(n,k-1)$ においてバックアップを取らないのが最適であっても,

$$-p(B+R)\leq d_n(k-1)$$

ならば, 状態 (n,k) においてバックアップを取るのが最適である.

3 パラメータが未知の場合

パラメータ p の値が未知の時, p がパラメータ s と t のベータ分布を, 事前分布としてもつと仮定できるものとする. 2節で定義された状態 (n,k) において, さらに p の事前分布がパラメータ (s,t) のベータ分布であるとき, 状態は (n,k,s,t) で表される. ここで, 観察値 $X=1$ を得た後の事後分布は, パラメータ $(s+1,t)$ のベータ分布である (たとえば[5] を参照).

状態(n,k,s,t)においてa_0を行うと，確率1で状態$(n-1,0,s,t)$に推移し，$k+1$個のファイルが安全に保存され，したがって$(k+1)R$の利得が生じる．ただし，このとき費用Cがかかる．また，状態(n,k,s,t)においてa_1を行うと，確率$s/(s+t)$で$k+1$個のファイルが失われ，状態$(n-1,0,s+1,t)$に推移する．このとき$(k+1)B$の損失が生じる．また，確率$t/(s+t)$ですべてのファイルは安全であり，このとき状態$(n-1,k+1,s,t+1)$に推移する．

今，$f_n(k;s,t)$を現在の状態が(n,k,s,t)のときの最大期待総利得とする．また，$i=0,1$に対して，$f_n^i(k;s,t)$は現在の状態が(n,k,s,t)のときに，まずa_iを行い，以後最適政策に従うときの最大期待総利得とする．このとき，$n=1,2,3,\cdots$および$k=0,1,2,\cdots$に対して，

$$f_n(k;s,t)=\max\left\{f_n^0(k;s,t),f_n^1(k;s,t)\right\} \tag{8}$$

および

$$f_0(k;s,t)=kR \tag{9}$$

が成立する．ここに，

$$f_n^0(k;s,t)=-C+(k+1)R+f_{n-1}(0;s,t) \tag{10}$$

および

$$f_n^1(k;s,t)=\frac{s}{s+t}\{-(k+1)B+f_{n-1}(0;s+1,t)\}+\frac{t}{s+t}f_{n-1}(k+1;s,t+1). \tag{11}$$

である．ここで，状態が$(n,k;s,t)$のときに，もし$f_n^i(k;s,t)\geq f_n^{1-i}(k;s,t)$ならば$a_i(i=0,1)$が最適であり，またもし$f_n^0(k;s,t)=f_n^1(k;s,t)$ならば，$a_0$と$a_1$の両方とも最適であるが，便宜上$f_n^0(k;s,t)=f_n^1(k;s,t)$ならば，$a_0$が最適であるという．このとき$n$についての帰納法により

$$-B\leq f_n(k;s,t)-f_n(k-1;s,t)\leq R. \tag{12}$$

および

$$-kB\leq f_n(k;s,t)-f_n(0;s,t)\leq kR. \tag{13}$$

が成立する．$n=1,2,3,\cdots$および$k=0,1,2,\cdots$に対して，

$$d_n(k;s,t)=f_n^0(k;s,t)-f_n^1(k;s,t)$$

とおくと

$$\begin{aligned}d_n(k;s,t)=&-C+(k+1)(R+\frac{s}{s+t}B)+f_{n-1}(0;s,t)\\&-\frac{s}{s+t}f_{n-1}(0;s+1,t)-\frac{t}{s+t}f_{n-1}(k+1;s,t+1).\end{aligned} \tag{14}$$

また

$$\begin{aligned}d_n(k-1;s,t)=&-C+k(R+\frac{s}{s+t}B)+f_{n-1}(0;s,t)\\&-\frac{s}{s+t}f_{n-1}(0;s+1,t)-\frac{t}{s+t}f_{n-1}(k;s,t+1).\end{aligned} \tag{15}$$

より

$$d_n(k;s,t)-d_n(k-1;s,t)=(R+\frac{s}{s+t}B)-\frac{t}{s+t}\{f_{n-1}(k+1;s,t+1)$$
$$-f_{n-1}(k;s,t+1)\}.$$

となる．ここで

$$f_{n-1}(k+1,s,t+1)-f_{n-1}(k,s,t+1)\leq R$$

により

$$d_n(k-1,s,t)+\frac{s}{s+t}(B+R)\leq d_n(k,s,t).$$

これより次の定理が導ける．

定理 2 (i) 状態$(n,k-1,s,t)$においてバックアップを取るのが最適ならば，状態(n,k,s,t)においてもバックアップを取るのが最適である．
(ii) 状態(n,k,s,t)においてバックアップを取るのが最適であっても，

$$\frac{s}{s+t}(B+R)\geq d_n(k,s,t)$$

ならば，状態$(n,k-1,s,t)$においてバックアップを取らないのが最適である．
(iii) 状態$(n,k-1,s,t)$においてバックアップを取らないのが最適であっても，

$$-\frac{s}{s+t}(B+R)\leq d_n(k-1,s,t)$$

ならば，状態(n,k,s,t)においてバックアップを取るのが最適である．

定理 3 $n=0,1,2,\cdots$ および $k=1,2,3,\cdots$に対して，

$$f_n(k;s+1,t)\leq f_n(k;s,t)\leq f_n(k;s,t+1)$$

証明．nについての帰納法による．$n=0$のときは明らかである．$n=1$のとき，

$$f_1^0(k;s+1,t)=f_1^0(k;s,t)$$

および

$$f_1^1(k;s+1,t)=f_1^1(k;s,t)-\frac{t}{(s+t)(s+t+1)}(k+1)(B+R),$$

であり，したがって

$$f_1(k;s+1,t)=\max\{f_1^0(k;s+1,t),f_1^1(k;s+1,t)\}$$
$$\leq\max\{f_1^0(k;s,t),f_1^1(k;s,t)\}$$
$$=f_1(k;s,t),$$

同様に

$$f_1(k;s,t)\leq f_1(k;s,t+1).$$

よって$n=1$のときも成立する．また

$$f_{n-1}(k;s+1,t) \le f_{n-1}(k;s,t) \le f_{n-1}(k;s,t+1).$$

が成立すると仮定する．このとき

$$f_n^0(k;s+1,t) - f_n^0(k;s,t) = f_{n-1}(0;s+1,t) - f_{n-1}(0;s,t)$$

であり，帰納法の仮定により

$$f_n^0(k;s+1,t) - f_n^0(k;s,t) \le 0 \tag{16}$$

また

$$\begin{aligned} f_n^1(k;s+1,t) - f_n^1(k;s,t) = &\left(\frac{s}{s+t} - \frac{s+1}{s+t+1}\right)(k+1)B \\ &+ \frac{s+1}{s+t+1} f_{n-1}(0;s+2,t) - \frac{s}{s+t} f_{n-1}(0;s+1,t) \\ &+ \frac{t}{s+t+1} f_{n-1}(k+1;s+1,t+1) - \frac{t}{s+t} f_{n-1}(k+1;s,t+1). \end{aligned}$$

ここで，帰納法の仮定により $f_{n-1}(0;s+2,t) \le f_{n-1}(0;s+1,t)$ および $f_{n-1}(k+1;s+1,t+1) \le f_{n-1}(k+1;s,t+1)$ だから

$$\begin{aligned} f_n^1(k;s+1,t) - f_n^1(k;s,t) \le \frac{t}{(s+t)(s+t+1)}\{&-(k+1)B + f_{n-1}(0;s+1,t) \\ &-f_{n-1}(k+1;s,t+1)\} \end{aligned}$$

さらに，$f_{n-1}(0;s+1,t) \le f_{n-1}(0;s,t)$，および $f_{n-1}(k+1;s,t+1) \ge f_{n-1}(k+1;s,t)$ により

$$\begin{aligned} f_n^1(k;s+1,t) - f_n^1(k;s,t) \le \frac{t}{(s+t)(s+t+1)}\{&-(k+1)B + f_{n-1}(0;s,t) \\ &-f_{n-1}(k+1;s,t)\} \end{aligned}$$

(12) により

$$f_n^1(k;s+1,t) - f_n^1(k;s,t) \le 0 \tag{17}$$

(16) および (17) より

$$f_n(k;s+1,t) - f_n(k;s,t) \le 0 \tag{18}$$

同様にして，

$$f_n(k;s,t) - f_n(k;s,t+1) \le 0 \tag{19}$$

が得られる．□

4 バンディット問題との比較

3節で述べたパラメータpが未知の場合のファイルの最適バックアップ問題は，要約すると次のようになる．

- 逐次的にn回の決定を行う．
- 各回においてa_0，a_1のいずれかの決定を行う．
- a_0を行うと，パラメータ$q=1$のベルヌーイ分布から観察値Xを得る．
- a_1を行うと，パラメータpのベルヌーイ分布から観察値Yを得る．
  $$P\{Y=1\}=1-P\{Y=0\}=p$$
- pの値は未知である．
- pはパラメータsとtのベータ分布を事前分布としてもつ．
- 状態は(n,k,s,t)で与えられる．
- 状態(n,k,s,t)においてa_0を行うと，費用Cがかかる．そして，確率1で状態$(n-1,0,s,t)$に推移する．このとき$k+1$個のファイルが安全に保存され，したがって$(k+1)R$の利得が生じる．
- 状態(n,k,s,t)においてa_1を行うと，
  - 確率$s/(s+t)$で状態$(n-1,0,s+1,t)$に推移する．このとき$k+1$個のファイルが失われ，したがって$(k+1)B$の損失が生じる．
  - 確率$t/(s+t)$で状態$(n-1,k+1,s,t+1)$に推移する．

これに対して，ベルヌーイ分布にしたがう2つの実験のうち，一方のパラメータが既知で，もう一方のパラメータが未知であり，未知パラメータの事前分布として，ベータ分布が仮定できる場合のバンディット問題を考える．このようなバンディット問題については，たとえば[1],[2],[8]等がある．このバンディット問題を要約すると次のようになる．

- 逐次的にn回の決定を行う．
- 各回においてa_0，a_1のいずれかの決定を行う．
- a_0を行うとパラメータqのベルヌーイ分布から観察値Xを得る．
  $$P\{X=1\}=1-P\{X=0\}=q$$
  このときの利得Xを得る．
- a_1を行うとパラメータpのベルヌーイ分布から観察値Yを得る．
  $$P\{Y=1\}=1-P\{Y=0\}=p$$
  このとき利得Yを得る．

- qの値は既知であるが，pの値は未知である．
- pはパラメータsとtのベータ分布を事前分布としてもつ．
- 状態は(n, s, t)で与えられる．
- 状態(n, s, t)においてa_0を行うと，
  - 確率qで$X = 1$を得て，状態$(n-1, s, t)$に推移する．
  - 確率$1-q$で$X = 0$を得て，状態$(n-1, s, t)$に推移する．
- 状態(n, s, t)においてa_1を行うと，
  - 確率$s/(s+t)$で$Y = 1$を得て，状態$(n-1, s+1, t)$に推移する．
  - 確率$t/(s+t)$で$Y = 0$を得て，状態$(n-1, s, t+1)$に推移する．

これら両者の違いは，前者の状態ベクトルの2番目のパラメータkの存在であり，これがこれらのモデルの本質的な違いであると考えられる．

参考文献

[1] Berry, D. A. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments.* Chapman and Hall.

[2] Bradt, R. N., Johnson, S. M. and Karlin, S. (1956). On sequential designs for maximizing the sum of n observations. *Annals of Mathematical Statistics* **27**, 1060-1074.

[3] Chandy, K. M., Browne, J. C., Dissly, C. W., and Uhrig, W. R. (1975). Analytic models for rollback and recovery strategies in data base systems. *IEEE Transactions on Software Engineering* **SE-1**, 100-110.

[4] Chandy, K. M. and Ramamoorthy, C. V. (1972). Rollback and recovery strategies for computer programs. *IEEE Transactions on Computer* **C-21**, 546-556.

[5] DeGroot, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

[6] Gelenbe, E. (1979). On the optimum checkpoint interval. *Journal of Association on Computing Machinary* **2**, 259-270.

[7] Kaio, N. and Osaki, S. (1985). A note on optimum check pointing policies. *Microelectron. Reliability* **25**, 451-453.

[8] Ross, S. M. (1983). *Introduction to Stochastic Dynamic Programming.* Academic Press, New York.

[9] Sandoh, H. and Kawai, H. (1991). An optimal N-job backup policy maximizing availability for a hard computer disk. *Journal of the Operations Research Society of Japan* **34**, 383-390.

[10] Toueg, S. and Babaoğlu, Ö. (1984). On the optimum checkpoint selection problem. *SIAM Journal of Computer* **13**, 630-649.

[11] Young, J. W. (1974). A first order approximation to the optimum checkpoint interval. *Communications of ACM* **17**, 530-531.